Data Visualization using Python and Streamlit

First A. Author, Diya Neupane, *Shaswat Pant, KU*

[1]

Abstract—**Data visualization plays a critical role in understanding complex datasets by transforming raw information into graphical representations that highlight patterns, trends, and insights. This project explores the use of Python—an accessible yet powerful programming language—along with Streamlit, an open-source framework, to create interactive and real-time data visualization applications. By leveraging libraries such as Pandas, Matplotlib, Seaborn, and Plotly within a Streamlit interface, users can effortlessly upload datasets, explore key metrics, and generate customizable visualizations through an intuitive web-based platform. The goal of this project is to simplify the visualization process for both technical and non-technical users, enabling efficient decision-making and exploratory data analysis through an engaging and dynamic user experience.**

I. INTRODUCTION

In today's data-driven landscape, the ability to quickly analyze and visualize datasets is crucial for informed decision-making across various domains. This report presents a comprehensive data visualization tool designed to streamline the process of exploratory data analysis for users working with CSV datasets. The tool is handy to overcome the common issue of mapping raw tabular data to useful visual insights without spending major technical skills or programming expertise.

Developed using Python and Streamlit, the main goal of this tool is to eliminate the barriers to data visualization and develop a user-friendly interface that automatically processes user-uploaded CSV files and generates appropriate visual representations based on data characteristics. This tool can help users coming from a variety of backgrounds to interpret datasets, removing the often intimidating and technical hurdles to deducing visualization.

II. METHODOLOGY

1. From Input CSV

An input module was created using Streamlit file uploader where users can browse and upload a CSV file that they want to visualize.

2. Attribute Classification

The uploaded CSV file was then read, and the type of attribute was determined to be either numeric or categorical using pandas. For the numeric data type, further, the boolean values (0 and 1) were checked, and the attribute was recognized as Categorical. Moreover, the numeric attributes with fewer than 20 unique values were termed as discrete, and the attributes with a number of unique values greater than or equal to 20 were termed as continuous.

Classification of the attributes into their types made it easier for the visualization logic to be implemented separately.

3. Visualization of a single attribute

The visualization of a single attribute was implemented through a Bar chart, a Histogram, a Line chart, and Pie chart (for categorical attributes). This was done using plotly.express, and the binning process was handled with a function.

4. Visualization of two attributes

Visualization of two attributes was handled using Scatter plot and Grouped bar chart which was also implemented using plotly.express. Users are able to select the attributes that they want placed on X and Y axes and view the relationships between the selected attributes accordingly.

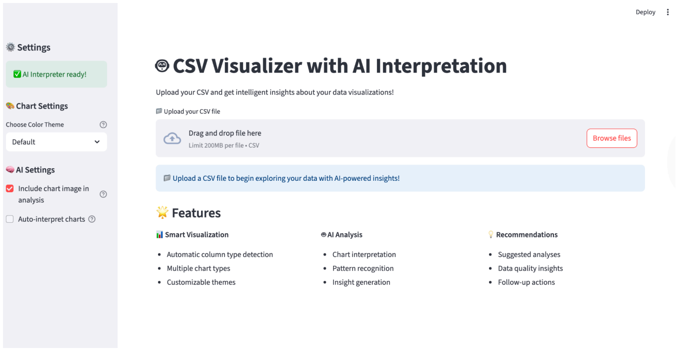
5. Interpretation using AI

Gemini API was used and a pipeline was created where, when a user clicks ‘Interpret with AI’ button, the just generated graph is passed to Gemini in order to generate a summary. Moreover, users are provided with a choice to provide a short description of the data being visualized which would give better interpretations.

[1]

III. FEATURES

1. Users can upload the CSV file they want to visualize.
2. Users can view the type(numeric(continuous/discrete) / categorical) of each attribute in the data and view missing values of each attribute.
3. Users can view Bar graphs, Histogram, Line graph, and Pie chart for selected singular attributes.
4. Users can change the bin size for the continuous numeric data.
5. Users can change the color palette of the charts.
6. Users can viewa Scatter Plot and a Grouped bar graph when two attributes are selected, where they can control what goes on the X axis and what goes on the Y axis.
7. Users can use AI to get an interpretation of the graph generated.



Deploy

Settings

✅ AI Interpreter ready!

Chart Settings

Choose Color Theme: Default

AI Settings

☑ Include chart image in analysis

□ Auto-interpret charts

CSV Visualizer with AI Interpretation

Upload your CSV and get intelligent insights about your data visualizations!

Upload your CSV file

Drag and drop file here
Limit 200MB per file • CSV
Browse files

Upload a CSV file to begin exploring your data with AI-powered insights!

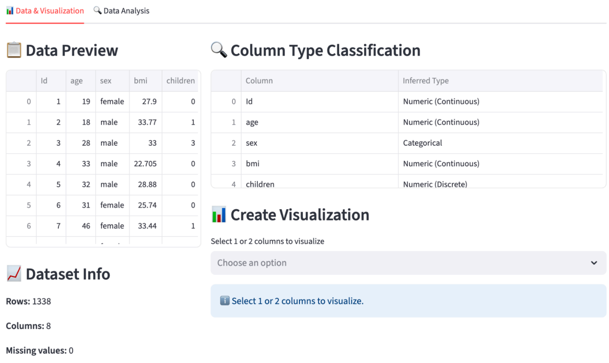
Features

Smart Visualization
- Automatic column type detection
- Multiple chart types
- Customizable themes

AI Analysis
- Chart interpretation
- Pattern recognition
- Insight generation

Recommendations
- Suggested analyses
- Data quality insights
- Follow-up actions

Fig.1: Landing UI of the platform



Data & Visualization | Data Analysis

Data Preview

	Id	age	sex	bmi	children
0	1	19	female	27.9	0
1	2	18	male	33.77	1
2	3	28	male	33	3
3	4	33	male	22.705	0
4	5	32	male	28.88	0
5	6	31	female	25.74	0
6	7	46	female	33.44	1

Column Type Classification

	Column	Inferred Type
0	Id	Numeric (Continuous)
1	age	Numeric (Continuous)
2	sex	Categorical
3	bmi	Numeric (Continuous)
4	children	Numeric (Discrete)

Dataset Info

Rows: 1338

Columns: 8

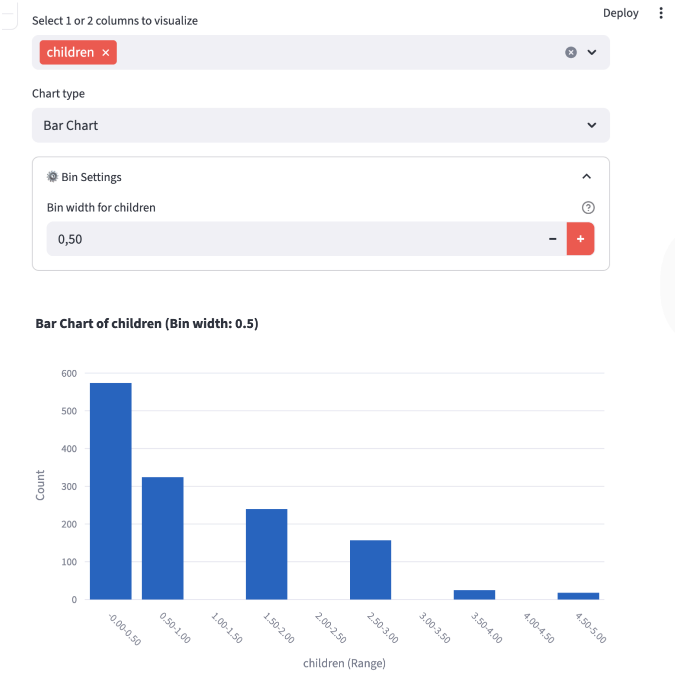
Missing values: 0

Create Visualization

Select 1 or 2 columns to visualize: Choose an option

Select 1 or 2 columns to visualize.

Fig.2: Data Preview



Deploy

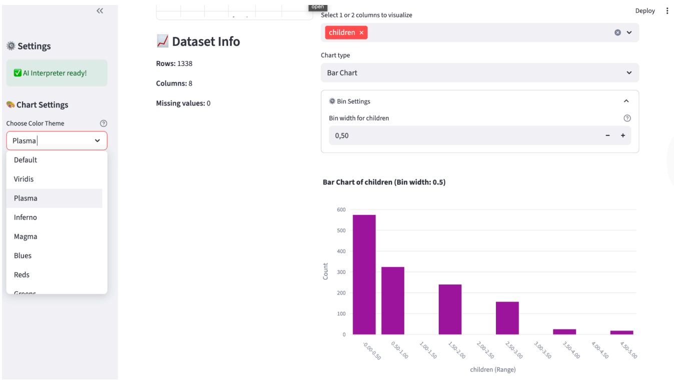
Select 1 or 2 columns to visualize: children

Chart type: Bar Chart

Bin Settings

Bin width for children: 0,50

Fig.3: Bar graph with bin width adjustment for one attribute



Settings

✅ AI Interpreter ready!

Chart Settings

Choose Color Theme: Plasma

Dataset Info

Rows: 1338

Columns: 8

Missing values: 0

Deploy

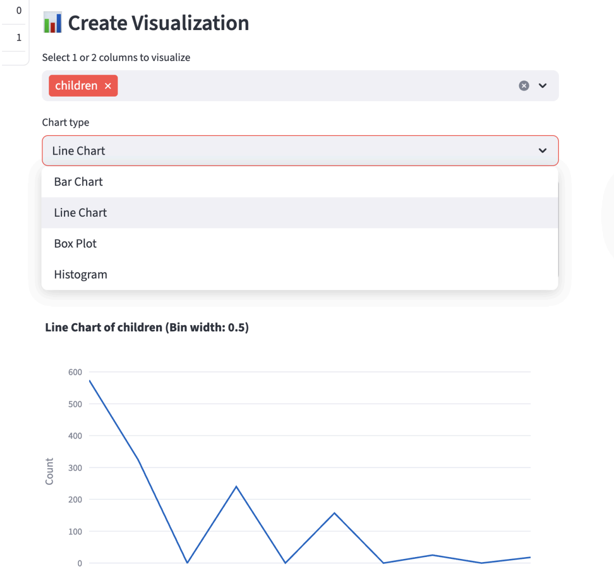
Select 1 or 2 columns to visualize: children

Chart type: Bar Chart

Bin Settings

Bin width for children: 0,50

Fig.4: Custom color palette choosing

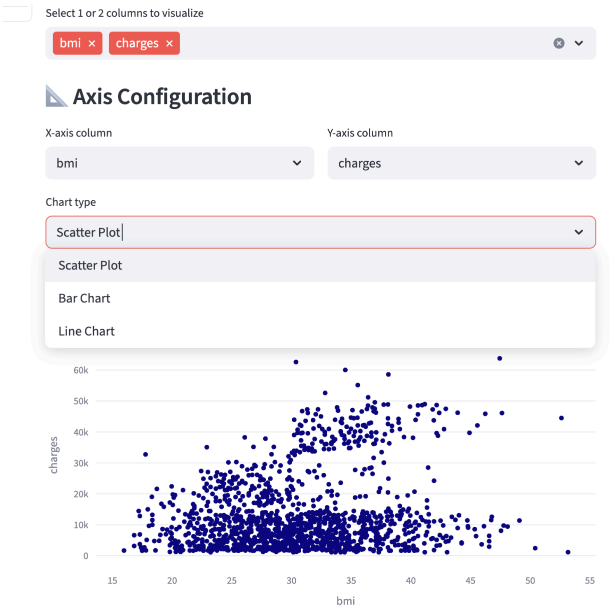


Create Visualization

Select 1 or 2 columns to visualize: children

Chart type: Line Chart

Fig.5: Options for charts



Select 1 or 2 columns to visualize: bmi × charges ×

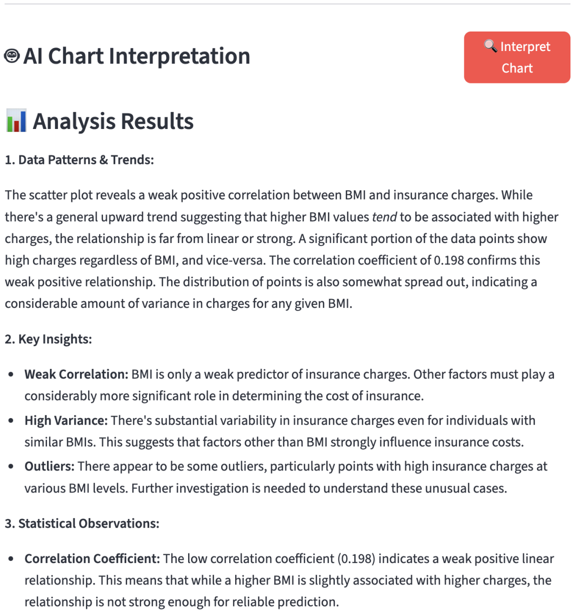
📐 Axis Configuration

X-axis column: bmi

Y-axis column: charges

Chart type: Scatter Plot

Fig.6: Options for visualization of multivariate data



AI Chart Interpretation

📊 Analysis Results

1. Data Patterns & Trends:

The scatter plot reveals a weak positive correlation between BMI and insurance charges. While there's a general upward trend suggesting that higher BMI values *tend* to be associated with higher charges, the relationship is far from linear or strong. A significant portion of the data points show high charges regardless of BMI, and vice-versa. The correlation coefficient of 0.198 confirms this weak positive relationship. The distribution of points is also somewhat spread out, indicating a considerable amount of variance in charges for any given BMI.

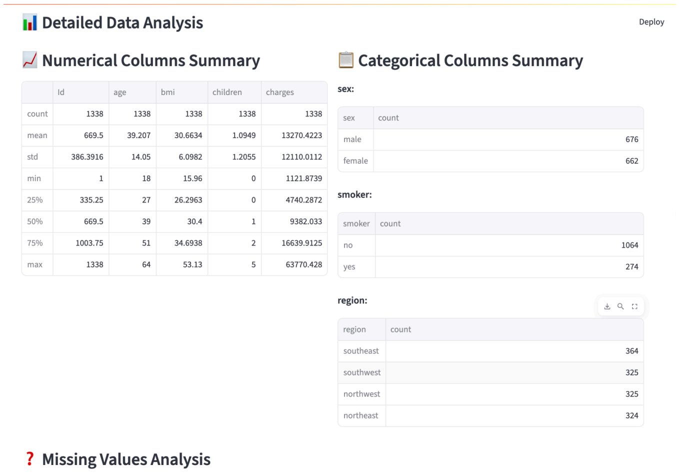
2. Key Insights:

- **Weak Correlation:** BMI is only a weak predictor of insurance charges. Other factors must play a considerably more significant role in determining the cost of insurance.
- **High Variance:** There's substantial variability in insurance charges even for individuals with similar BMIs. This suggests that factors other than BMI strongly influence insurance costs.
- **Outliers:** There appear to be some outliers, particularly points with high insurance charges at various BMI levels. Further investigation is needed to understand these unusual cases.

3. Statistical Observations:

- **Correlation Coefficient:** The low correlation coefficient (0.198) indicates a weak positive linear relationship. This means that while a higher BMI is slightly associated with higher charges, the relationship is not strong enough for reliable prediction.

Fig.7: AI interpretation of charts



📊 Detailed Data Analysis

📈 Numerical Columns Summary

	Id	age	bmi	children	charges
count	1338	1338	1338	1338	1338
mean	669.5	39.207	30.6634	1.0949	13270.4223
std	386.3916	14.05	6.0982	1.2055	12110.0112
min	1	18	15.96	0	1121.8739
25%	335.25	27	26.2963	0	4740.2872
50%	669.5	39	30.4	1	9382.033
75%	1003.75	51	34.6938	2	16639.9125
max	1338	64	53.13	5	63770.428

📋 Categorical Columns Summary

sex:

sex	count
male	676
female	662

smoker:

smoker	count
no	1064
yes	274

region:

region	count
southeast	364
southwest	325
northwest	325
northeast	324

❓ Missing Values Analysis

Fig.8: Data analysis features including numeric summary and missing values of each attribute

IV CONCLUSION

This project successfully delivers a robust and user-friendly data visualization tool that simplifies exploratory data analysis for users of all backgrounds. By leveraging Python, Streamlit, and Plotly, the system automates the detection of attribute types and generates intuitive visualizations tailored to the nature of the data. The seamless integration of AI-powered interpretation through the Gemini API further enhances the analytical value by offering context-aware insights into the visualized data.